Chemical forensic approach for production batch classification: a methodological study

G. Delaporte[1], M. Spiandore[1], C. Albaret[1], E. Joubert[1]

[1] DGA Maîtrise NRBC, Département d'analyse chimique, 91710 Vert-Le-Petit, France,
gregoire.delaporte@intradef.gouv.fr

Keywords: Chemical Forensic, Classification, GC-MS.

1 Introduction

Chemical forensic is an emerging field that "aims to provide information to support the attribution of a chemical (or mixture) of interest to its source" [1]. It is of special interest in the field of chemical weapon use allegation attribution.

The general methodology of chemical forensics has already been described in several articles in the past few years [2-3]. Scaling up this approach from academic research to practical implementation rises many challenges. It would indeed require the contributions from several laboratories, and the management of an important data variability (different instrument set-ups, complex matrices, several representative sources of different processes or synthetic routes, manufacturer, stock or batch, localization etc.). Due to those constraints, an in-deep methodological study is prerequisite to push chemical forensic approaches to a more operational level.

A first data set (data set n°1) has been used for a feasibility study of classification of methyphosphonic dichloride (DC) samples according to their production batches by GC-MS. This preliminary study aims at validate the data analysis strategy.

Then, a methodological study have been conducted on the first step of the data analysis process: raw data preprocessing. It aims is to explore the behavior of common raw MS data preprocessing algorithms (namely XCMS and MZmine) on two data sets:

- Data set n°2: several standard solutions with compounds exhibiting various concentration and chromatographic behavior (peak shapes, coelutions etc.)
- Data set n°3: Standard solutions and a nerve-agent precursor (DC) analyzed in 5 different laboratories worldwide

Together they are the first building blocks of a more comprehensive study on the implementation of chemical forensic approach in "real-life" cases.

2 Material and methods

All data analyzed have been acquired on GC-MS systems using various instrumental set-ups.

Data set n°1 consists in DC samples from 4 different production batches which were treated according to the procedure described in [2] and analyzed on a GC-MS system using a SolGelWax column (SGE, 30m x 0,25 mm x 25 µm), EI ionization and quadrupole mass analyzer. All sample were analyzed in at least 3 technical replicates, each one analyzed thrice.

Data set n°2 consists in six standard solutions (custom laboratory quality control solution at three dilution levels -0.1 → 10 µg/mL-; custom sulfur compound mix; commercial phosphorous flame retardant mix) analyzed on the same GC-MS system as data set n°1 but with a VF5-ms column

(Agilent J&W, 30m x 0,25 mm x 0.25 µm). Compounds were chosen for their diversity of GC-MS responses both in terms signal intensity and chromatographic peak shapes.

Data set n°3 consists in a standard solution (custom standard solution at 10 µg/mL) and a DC sample analyzed in 5 different laboratories with different instrument set ups, all involving a SolGelWax column (various suppliers, 30m x 0,25 mm x 25 µm), an EI ion source and a quadrupole mass analyzer.

All the data have been processed using MZmine 2.53 [4], XCMS [5-7], R (4.0.3) and Matlab (R2020a).

3 Results and discussion

The feasibility study shows that the developed data analysis workflow is suitable for the classification of DC samples according to their production batches with satisfactory results.

Methodological studies of raw GC-MS data preprocessing (data set n°1 & 2) give the following outcome:

- Commonly used raw data preprocessing tools, commonly used in metabolomics (MZmine and XCMS) are able to handle the large data variability, once appropriate parameters are set after a careful data visualization and some optimization;
- Peak integration values are very similar to the reference ones obtained by a manual data review, even for highly complex samples.

4 Conclusion

This study is a first step towards the implementation of chemical forensics approach on a larger scale.

The global data analysis workflow seems suitable for sample classification according to their impurity profile.

The generic raw MS data preprocessing tools, commonly used in the field of metabolomics, are able to handle raw data exhibiting high variability, both by detecting all the relevant compounds and by giving peak integration close to the reference ones.

Further work would imply the study of the behavior of commonly used classification methods (such as PLS-DA for example) on data sets containing a high level of variability (different GC-MS instrumental set-ups, different laboratories etc.), such as the ones encountered in chemical forensics.

5 References

[1] C. G. Fraga, Chemical forensics, *Talanta*, vol. 186, p. 585, 2018.

[2] C. G. Fraga, G. A. Pérez Acosta, M. D. Crenshaw, K. Wallace, G. M. Mong, et H. A. Colburn, Impurity Profiling to Match a Nerve Agent to Its Precursor Source for Chemical Forensics Applications, *Analytical Chemistry*, vol. 83, p. 9564-9572, 2011.

[3] K. H. Holmgren et al., Synthesis route attribution of sulfur mustard by multivariate data analysis of chemical signatures, *Talanta*, vol. 186, p. 615-621, 2018.

[4] T. Pluskal, et al., MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data, BMC Bioinformatics 11:395 (2010). PMID: 20650010

[5] Smith, C.A., Want, E.J., O'Maille, G., Abagyan,R., Siuzdak, G. (2006). "XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching and identification." Analytical Chemistry, 78, 779–787.

[6] Tautenhahn R, Boettcher C, Neumann S (2008). "Highly sensitive feature detection for high resolution LC/MS." BMC Bioinformatics, 9, 504.

[7] Benton HP, Want EJ, Ebbels TMD (2010). "Correction of mass calibration gaps in liquid chromatography-mass spectrometry metabolomics data." BIOINFORMATICS, 26, 2488.