Potentials and Challenges of Data Processing Strategies in Non-Target Analysis of Water Samples

Lotta Laura Hohrenk-Danzouma (lotta.hohrenk-danzouma@leuphana.de)

Abstract

Non-target screening (NTS) is an emerging field for environmental monitoring. It enables a more comprehensive overview of water pollution and the discovery of unknown substances and opens the way to new ways of data analysis. However, data processing requires sophisticated strategies to achieve meaningful results and still faces several challenges. The consistency of feature extraction with different programs was assessed and showed the need for more robust methods and harmonized quality control criteria. The benefits of implementing complementary multivariate chemometric tools for feature extraction and prioritization were highlighted. Future advancements, including AI and integration with other data forms, promise to enhance NTS capabilities.

Introduction

Non-target screening (NTS) and suspect screening are becoming increasingly important tools for environmental monitoring and evaluation of wastewater or drinking water treatment processes. This technique is based on high-resolution mass spectrometry (HRMS) coupled to liquid chromatography (LC) and offers the potential for a selective and sensitive detection of a wide range of organic micropollutants (OMPs) at trace concentrations within a single measurement. As this method is not limited to pre-defined analytes, it provides a more holistic picture of the contaminant load in the aquatic environment and allows the discovery of previously unknown compounds such as transformation products [1]. However, with HRMS-based screening methods large amounts of data are recorded and sophisticated data processing strategies are necessary. Data processing and mining is often not only the most time-consuming part but also a crucial step to obtain meaningful results and to be able to exploit the full potential of NTS. Thus, the development of feature extraction algorithms, prioritization strategies, and identification approaches has become a major development field in the NTS area [2].

There are many different software tools and programs available for processing NTS data and the procedure needs to be adapted depending on the research question. Nevertheless, some general steps can be summarized and are shown in Figure 1. The two main tasks are: data *pre-processing* or *feature extraction* and data *post-processing* or *data analysis*.

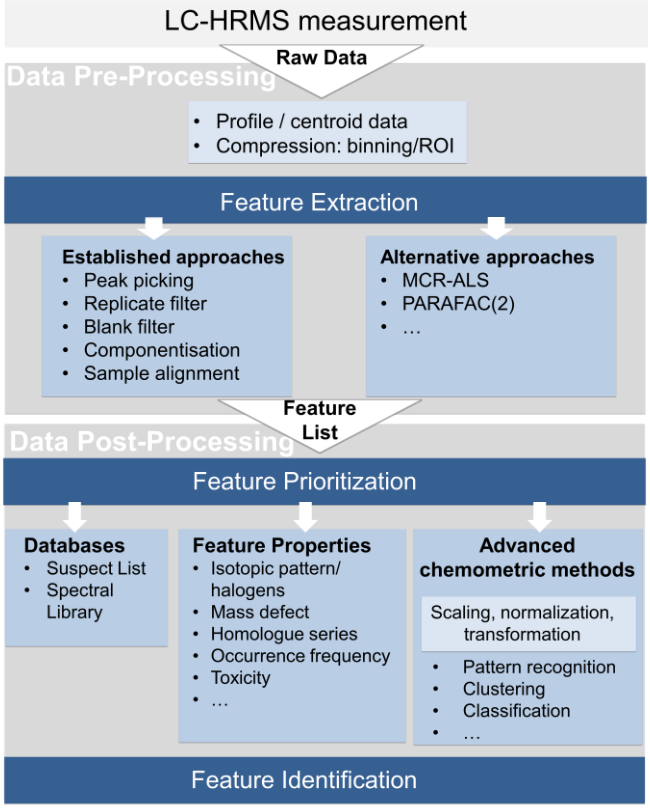

Figure 1: General data processing workflow for NTS: from raw data to feature extraction and prioritization to final feature identification.

The purpose of data pre-processing is a reliable extraction of analyte signals from raw data. Analyte signals are commonly referred to as features and are defined by their m/z, retention time and intensity. Any software or algorithm used for feature extraction has to deal with the challenge of distinguishing between real analyte signals and noise. Especially for low abundant signals, this can be challenging. Therefore, multiple steps are usually required to filter out false positives without losing true positives. Even after various filter steps, complex datasets with usually thousands of features remain [3]. Therefore, it is subsequently essential to reduce and prioritize features that are relevant to the studied research question, and which can later be identified. The range of possible prioritization strategies is just as diverse as the range of possible applications of NTS. In many cases, features with high intensity, occurrence frequency or certain properties (e.g. presence of halogens, mass defect,…) are selected. Or, with the approach known as suspect screening, m/z values of compounds of particular interest as e.g. transformation products or new contaminant classes of concern can be systematically searched for. In addition, several chemometric data mining strategies can be employed to evaluate differentiation/similarity between samples and discover hidden pollution patterns and trends.

In this article several aspects of data processing strategies, focusing on both the feature extraction step and feature prioritization step based on multivariate chemometric methods will be discussed. The importance of the feature extraction step is emphasized by first identifying weaknesses in the consistency of results obtained from different programs and secondly presenting an alternative chemometric-based feature extraction approach. Finally, the benefits of feature prioritization based on multiple complementary multivariate chemometric methods for NTS data were highlighted and future trends are outlined.

Comparison of feature extraction tools

The comparability of feature extraction with four different commonly used open-source and commercial software tools (MZmine2, enviMass, Compound Discoverer, XCMS online) was examined in a first publication [4]. For this purpose, feature extraction with each software tool was performed on the same raw data set and the overlap of resulting feature lists was analysed. Results show a low coherence between different processing tools, as the overlap of features between all four programs was around 10%, and for each software between 40% and 55% of features did not match with any other program. The deviating implementation of filtering steps such as replicate- and blank filter was identified as one source of observed discrepancies. This comparison showed the necessity for higher robustness of data processing tools, a better understanding of algorithms as well as the influence of different parameter settings for each approach. Even though a general standardization of feature extraction is not feasible, a higher awareness of the impact of this step and a transparent and detailed reporting of the entire data processing workflow were encouraged with this work.

Alternative feature extraction approaches – MCR-ALS

In addition, an alternative feature extraction procedure based on chemometric models such as regions of interest (ROI) and multivariate curve resolution alternating least squares (MCR-ALS) was employed on an NTS dataset of water samples for the first time [5]. This approach circumvents several error-prone processing steps as there is no need for chromategraphic alignment or grouping of multiple features of the same analyte. In a nutshell: This approach is based on the premise that the measured variation in all samples can be described by a set of "MCR-ALS components“. Each MCR-ALS component represents an OMP, however, unlike a feature which is limited to one retention time and m/z value, they combine a resolved pure mass spectra and elution profiles for each OMP signal. By that, later filtering steps for adducts or isotope peaks are not necessary. The process consists of several steps, which are summarized and graphically illustrated in Figure 2. At first, the raw data are compressed based on ROI approach. To analyze different samples simultaneously, an augmented data matrix is created by stacking several data matrices on top of each other. Matrices are augmented based on common ROI, therefore retention time alignment or correction are not required. After these data compression and augmentation steps, a set of MCR-ALS components of both chromatographic profiles and pure spectra is determined by alternating least-squares optimization approach starting by a set of initial estimates in an iterative process [6].

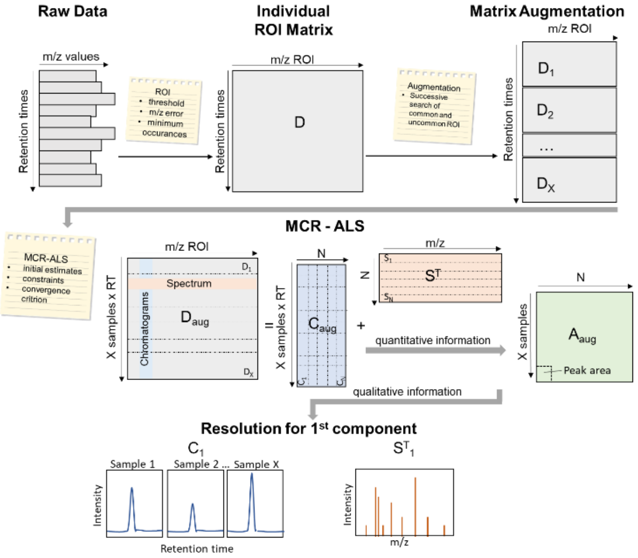

Figure 2: Overview of different steps of ROI/MCR-ALS processing pipeline: Raw data compression via ROI procedure and matrix augmentation followed by MCR-ALS resolution of components. Abbreviations: Aaug: matrix of peak areas of N components and X samples; Caug: augmented column vectors of the elution profiles of N components D: data matrix, Daug: augmented data matrices; N: number of components; ST: row vectors of pure spectra of N components; x: number of samples.

The approach was employed on samples with different complexity including a target data set of spiked drinking water samples and a NTS dataset obtained from different steps of a waste-water treatment plant and upstream of the receiving river. For all samples, chromatographic peaks and the corresponding mass spectra of OMPs were fully resolved in the presence of highly co-eluting irrelevant and interfering peaks. In the next step, features discriminating between several sample classes of the NTS data set were prioritized based on several multivariate and univariate chemometric methods. By that, from 101 resolved MCR-ALS components, 24 were selected and tentatively identified.

Chemometric feature prioritization

Apart from robust and reliable feature extraction, the prioritization of relevant features is crucial in NTS to extract the information of interest. The benefits of feature prioritization based on multiple complementary multivariate chemometric methods for NTS data were highlighted in a further publication [7]. Temporal and spatial trends on a data set were analyzed with different complementary unsupervised (PCA: principal component analysis and HCA: hierarchical cluster analysis) and supervised (ASCA: ANOVA simultaneous component analysis [8] and PLS-DA: partial least squares discriminant analysis [9]) chemometric approaches. The contribution of each approach to an overall deeper understanding of samples and hidden pollution patterns and to find a subset of discriminating features between samples was illustrated. Samples were obtained from a passive sampler monitoring campaign of three small streams and one major river (spatial factor) over four sampling periods (time factor). Unsupervised explorative chemometric tools were used to obtain a general overview of samples, where mainly spatial differences were visible. Subsequently, the ASCA approach was used to obtain deeper insights to the data set and disentangle the influence of spatial and seasonal effects as well as their interaction. The workflow was applied on a target and non-target dataset that both showed a dominant influence of different sampling locations and individual temporal pollution patterns for each river. With the limited set of target analytes, general seasonal pollution patterns were apparent, but NTS data provide a more holistic view of site-specific pollutant loads. With a complementary partial least squares-discriminant analysis (PLS-DA) and Volcano-based prioritization strategy, 223 site and 45 season-specific features were selected and tentatively identified. However, the majority of features that appear as relevant with this approach could not be identified ultimately. This shows that unknown identification remains a bottleneck in NTS data processing. Thus, there is a need to extend databases and develop new techniques for unknown identification. The presented workflow can be transferred to many other environmental datasets and different research questions including combined spatial and temporal investigations.

Conclusion and Outlook

Overall, it was demonstrated that data processing is crucial in NTS to obtain meaningful results for comprehensive environmental monitoring. For both feature extraction as well as prioritization remaining challenges and the capabilities of the implementation of advanced multivariate chemometric tools were highlighted. In the future exciting developments in the field of NTS data processing are expected. The rapid development of artificial intelligence programs can be an accelerator in solving the remaining data processing challenges at all stages of the workflow. For example, there is a need for feature extraction that is not only more robust but also more automated, so that it can be applied in less time for widespread practical use of NTS methodology. This also applies to data mining tools, which have so far only been used to a limited extent as they require extensive programming skills. Other promising developments in this area are the coupling of NTS data with other methods, such as effect-driven analysis (EDA) or other “omics” data such as metagenomics, transcriptomics [10] etc. With these advancements, the full potential of NTS as a tool to complete our knowledge and understanding of environmental pollution patterns, risks and remediation strategies can be achieved.

References

[1] Schmidt, T. C. Analytical and Bioanalytical Chemistry 2018, DOI: 10.1007/s00216-018-1015-9.

[2] Hollender, J.; Schymanski, E. L.; Singer, H. P.; Ferguson, P. L. Environmental Science & Technology 2017, DOI: 10.1021/acs.est.7b02184.

[3] Katajamaa, M.; Oresic, M. Journal of Chromatography. A 2007, DOI: 10.1016/j.chroma.2007.04.021.

[4] Hohrenk, L. L.; Itzel, F.; Baetz, N.; Tuerk, J.; Vosough, M.; Schmidt, T. C. Analytical Chemistry 2020, DOI: 10.1021/acs.analchem.9b04095.

[5] Hohrenk, L. L.; Vosough, M.; Schmidt, T. C. Analytical Chemistry 2019, DOI: 10.1021/acs.analchem.9b01984.

[6] Gorrochategui, E.; Jaumot, J.; Tauler, R. BMC Bioinformatics 2019, DOI: 10.1186/s12859-019-2848-8.

[7] Hohrenk-Danzouma, L. L.; Vosough, M.; Merkus, V. I.; Drees, F.; Schmidt, T. C. Environmental Science & Technology 2022, DOI: 10.1021/acs.est.1c08014.

[8] Smilde, A. K.; Jansen, J. J.; Hoefsloot, H. C. J.; Lamers, R.-J. A. N.; van der Greef, J.; Timmerman, M. E. Bioinformatics (Oxford, England) 2005, DOI: 10.1093/bioinformatics/bti476.

[9] Ballabio, D.; Consonni, V. Analytical Methods 2013, DOI: 10.1039/c3ay40582f.

[10] Sieber, G.; Drees, F.; Shah, M.; Stach, T. L.; Hohrenk-Danzouma, L.; Bock, C.; Vosough, M.; Schumann, M.; Sures, B.; Probst, A. J.; Schmidt, T. C.; Beisser, D.; Boenigk, J. The Science of the Total Environment 2023, DOI: 10.1016/j.scitotenv.2023.167457.

Author

Dr. Lotta Hohrenk-Danzouma
Institut für Nachhaltige Chemie, Fakultät Nachhaltigkeit
Leuphana Universität Lüneburg
Universitätsallee 1
21335 Lüneburg
Telefon: 04131 6771350
E-Mail: lotta.hohrenk-danzouma@leuphana.de